Bioinformaticsとソフトウエア (符号と暗号の代数的数理)

AUTHOR(S):

田辺, 隆人

CITATION:

田辺, 隆人. Bioinformaticsとソフトウエア (符号と暗号の代数的数理). 数理解析研究所講究録 2005, 1420: 28-40

ISSUE DATE:

2005-04

URL:

http://hdl.handle.net/2433/47179

数理解析研究所講究録 1420 巻 2005 年 28-40

28

Bioinformatics とソフトウエア

(株)数理システム　田辺隆人(Takahito Tanabe)
Mathematical Systems Inc.

Cypripedium

Lady's slipper orchid

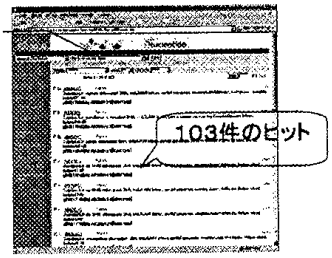
主要な配列データベース

- GenBank(**http://www.ncbi.nlm.nih.gov/**)
- EMBL (**http://www.ebi.ac.uk/embl/**)
- DDBJ (**http://www.ddbj.nig.ac.jp/**)
- PIR (**http://www-nbrf.georgetown.edu/**)
- EXPASY (**http://www.expasy.ch/**)
- SRS (**http://srs6.ebi.ac.uk/**)

データベース検索

- NCBI, ヌクレオチドデータベース

"Cypripedium Japan" で検索



データベース検索

- 個々のアイテムの概要

country:Canada:Prince Edward Island, isolate:AC-02
gi|13517201|dbj|AB056315.1|[13517201]

4. AB056314 Reports
Cypripedium sp. U-03 chloroplast DNA, trnL(UAA) intron, partial sequence,
country:Japan:Hokkaido, Rebun island, isolate:U-03
gi|13517200|dbj|AB056314.1|[13517200]

5. AB056313 Reports
Cypripedium sp. U-02 chloroplast DNA, trnL(UAA) intron, partial sequence,

データベース検索

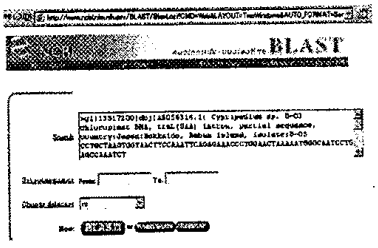
- 配列(FASTA形式)

4. AB056314 Reports Cypripedium sp. U...
[gi|13517200]

>gi|13517200|dbj|AB056314.1| Cypripedium sp. U-03 chloroplast DNA, trnL(UAA) intron, partial
CCTGCTAAGTGCTAACTTCCAAATTCAGACAAACCCTCGAACTAAAAATCGGCAATCCTGAGCCAAATCT
TTGTTTTTATAAAAAATGCAAAATTAGAATAAAAAGGGATAGGTGCAGACACTCAATGGAACCTGTTCTA
ACGAATCAAATTGACTACTTAAAATGGAAGACTTATTCTCAATCCATTCCAATTGAAGTTGAAAAAGGAA
TAGAATTCAAATATTCAGTGATCAAATCAAATGATTCATTCCAGAGTTTCATAGATCTTTTGAAAGTTAA
AAAGTTAATCGCAGCGAGAATAAACAGAGAGCTCCCATTTTACATGTCAATATCCACAACAATGAAA

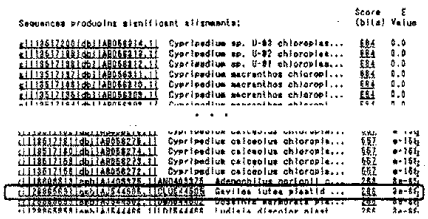
アラインメント

- BLAST(類似の配列の検索)

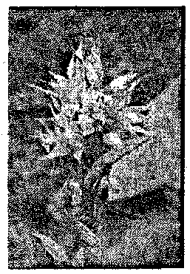


アラインメント

- BLAST結果（スコア順）



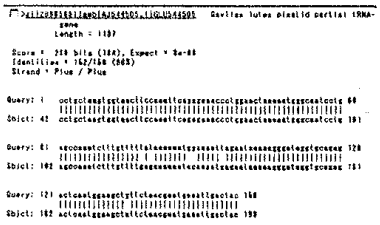
Gavilea lutea



Torres Del Paine, Patagonia, Chile, South America
January 2003
by Tom Kornack

アラインメント

- 個々の結果



アラインメント

- 「ぶれ」の許容

```
tttataaaaaatggaaaatta
||  | |||||||  |||||  |
ttgagaaaaaatacaaaatga
```

アラインメント

- 計算機科学の古典的問題

```
facet% diff file1 file2  #ファイルの違いを探すUNIXツール
5c5
<   ifstream infile(argv[1]);
---
>   ifstream myfile(argv[1]);
10c10
<   while( infile >> val ) {
---
>   while( myfile >> val ) {
facet%
```

大域的アラインメント

Needleman-Wunschアルゴリズム（1970）

- 両方をすべて使う一致を見つける
  1. 格子の左上から最大スコア値Dを計算（漸化式を適用）
  2. 右下のDが最大スコアの厳密解
  3. 右下から最大スコアを与える選択を逆に辿ってパスを発見

動的計画法

漸化式

Needleman-Wunschアルゴリズム

$$F(i,j)=\max\begin{cases}F(i-1,j-1)+s(x_i,y_j)\\ F(i-1,j)-d\\ F(i,j-1)-d\end{cases}$$

一致するときのスコア

どちらかを飛ばしたときのペナルティ

$$F(i,0)=-id$$
$$F(0,j)=-jd$$

開始点をずらしたときのペナルティ

インパクト

- 進化の経路の解析

シロイヌナズナ

Arabidopsis thaliana

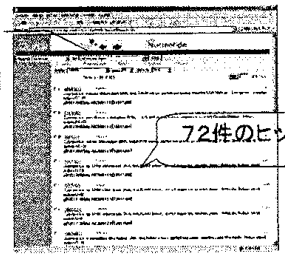
シロイヌナズナは、世代時間が短くゲノムがコンパクトなため遺伝学のモデル植物として広く使われ、多数の突然変異、DNAクローン、形質転換系統等の資源が利用可能です。セントロメアやテロメアやrDNAの反復配列領域を除いて、ゲノムの全塩基配列がほぼ決まっています。ゲノムDNAがメチル化されている真核生物の中で最も低ゲノム冗長配列が決まったのはシロイヌナズナです（酵母やC.elegansではゲノムDNAメチル化が見つかりません）。他の生物は植物生理現象の研究に使われていますが、脊椎動物にまで共通するエピジェネティックな現象の研究にも適した材料です。

アラインメント解析で69%の遺伝子機能が判明
（実験的には9%）

データベース検索

- NCBI, 構造データベース

Chloroplastで検索



72件のヒット

データベース検索

- 個々の結果

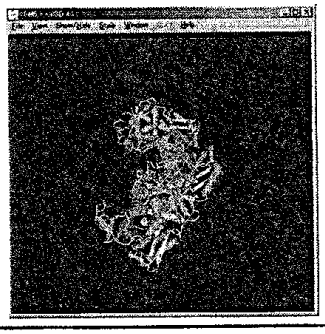
1: 1RM5
Crystal Structure Of Mutant S188a Of Photosynthetic Glyceraldehyde-3-Phosphate Dehydrogenase A4 Isoform, Complexed With Nadp
[mmdbId:28548]

Description: Crystal Structure Of Mutant S188a Of Photosynthetic Glyceraldehyde-3-Phosphate Dehydrogenase A4 Isoform, Complexed With Nadp
Deposition: F.Sparla, S.Fermani, G.Falini, A.Ripamonti, P.Sabatino, P.Pupillo & P. Trost, 27-Nov-03
Taxonomy: Spinacia oleracea
Reference: PubMed **MMDB:** 28548 **PDB:** 1RM5

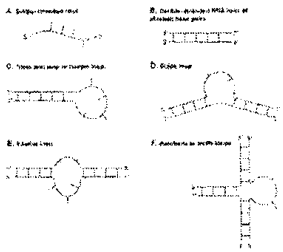
ほうれん草

データベース検索

- 三次構造

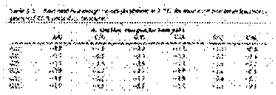


二次構造特定
RNAfolding

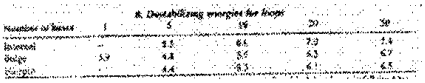


二次構造特定
入力データ

- RNA塩基配列
- 二塩基間結合のエネルギー

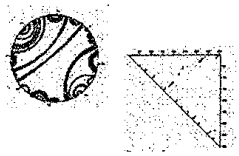


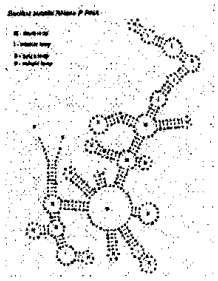
- 各種ループ構造のエネルギー



二次構造特定
アルゴリズム

- エネルギー最小化





二次構造特定
アルゴリズム

$$W(i,j) = \min\{W(i+1,j), W(i,j-1), V(i,j), \min_{i \le k < j} W(i,k) + W(k+1,j)\} \quad (1)$$

and

$$V(i,j) = \min\{eh(i,j), es(i,j) + V(i+1,j-1), VBI(i,j), VM(i,j)\}, \quad (2)$$

where

$$VBI(i,j) = \min_{\substack{i<i'<j'<j \\ i'-i+j-j'>2}} \{eb(i,j,i',j') + V(i',j')\} \quad (3)$$

and

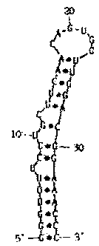
$$VM(i,j) = \min_{i<k<j-1} \{W(i+1,k) + W(k+1,j-1)\}. \quad (4)$$

The minimum folding energy, E_{min} is given by $W(1,n)$

エネルギー表の値

2-loopまで考慮

二次構造の特定
mfoldサーバー



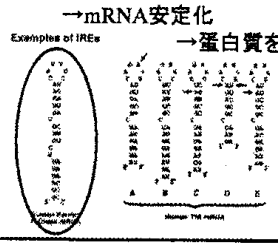
"GGGUUUUCCUGCUUCAACAG UGCUUGGACGGAAACCC"に対する mfoldサーバーの応答

http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi

dG = -10.1 [initially -10.1] ferritin

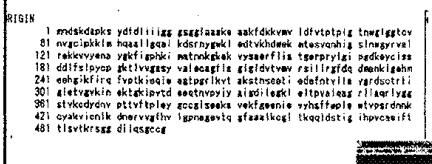
インパクト
RNA機能調整の解明

- 血中フェリチン
  - IRE(28ゲノム)にIRE結合蛋白が結合
    →mRNA安定化
    →蛋白質を大量に合成



特徴的な構造

三次構造特定
入出力


ORIGIN
1 emdskdqpks ydtdliiigg esggiaasko aakfdkkvmv ldfvtelpig tnegiggtcv
61 nvgcipkkln hqaallgcwi kdcrnygwki edtvkhdmnk etscvqnhig slnegyrval
121 rakkcvyvna ygkfigpkki matnskgkak vysaerflia tgerpryigi sqdkcvycis
181 ddlfslpycp gktlvvgasy valeaagtla gigidvtvav rsillrgfdq dmankigahn
241 eehgikfirq fvptkiegle egtpgrlkvt akstnseeti edafdtvila vgrdscrtri
301 gletvgvkin ektgkipvtd eeetnvpyiy aigdilegkl eltpvaleag rllaerlygg
361 stvkcdydnv pttvftpley gccglseeks vekfgeenie vyhsffqple mtvpsrdnnk
421 cyakvicnlk dnervvgfhv lgpnagevtq gfaaalkcgl tkqqldstig ihpvcaeifl
481 tlsvtkrsgg dilqsgccg

三次構造特定

- コンペティション

Fifth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

Asilomar Conference Center
December 2002

三次構造特定 アルゴリズム

- 断片(3－9残基)の検索
- MonteCarlo法
- Superfamilyに特徴的な構造情報利用
- ポテンシャル関数(疎水性・free-energy. .)
- 相同解析の結果を利用
- Knowledge base 検索

帰納的な方法が不可欠

機械学習 ソフトウエア

- ソフトウエア
  HMMpro(隠れマルコフモデル解析ソフト)

  開発元:Net-ID Inc.
- 応用
  - 遺伝子発見
  - 構造情報の取得
  - 相同解析

隠れマルコフモデル

- 記号列
  - DNA/RNAアルファベット(4種)

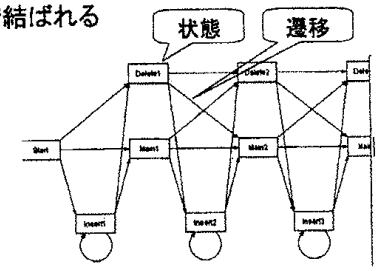
    A = Adenine　C = Cytosine
    G = Guanine　T = Thymine／U = Uracil
  - アミノ酸(20種)

    A = Alanine (Ala)　R = Arginine (Arg)
    D = Aspartic Acid (Asp)　C = Cysteine (Cys)
    Q = Glutamine (Gln)　E = Glutamic Acid (Glu)
    N = Asparagine (Asn)　G = Glycine (Gly)
    H = Histidine (His)　I = Isoleucine (Ile)
    L = Leucine (Leu)　K = Lysine (Lys)
    M = Methionine (Met)　F = Phenylalanine (Phe)
    P = Proline (Pro)　S = Serine (Ser)
    T = Threonine (Thr)　W = Tryptophan (Trp)
    Y = Tyrosine (Tyr)　V = Valine (Val)

状態遷移確率＋出力確率

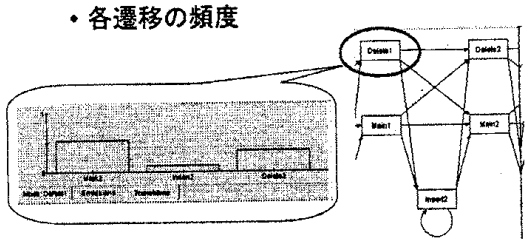
状態列

- 遷移で結ばれる



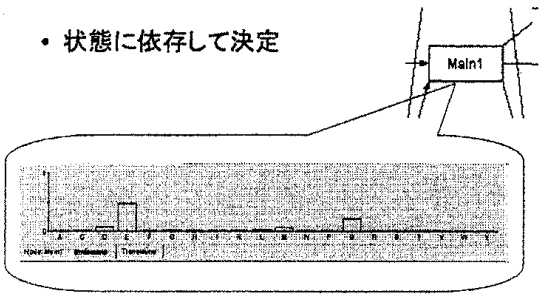
状態遷移確率

- 各遷移の頻度



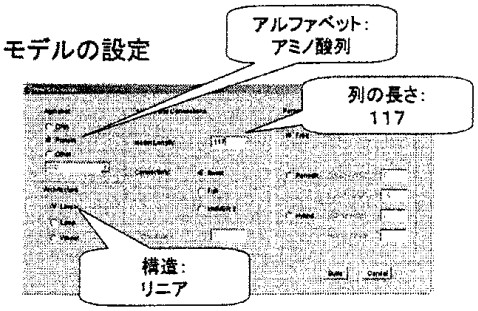
出力確率

- 状態に依存して決定



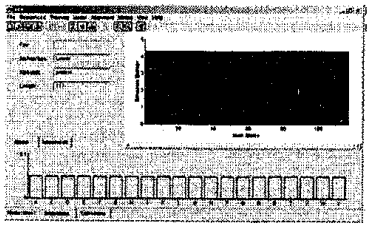
マルコフモデルの学習

- モデルの設定



学習前

- 出力分布



学習元ファイルのインポート

- ファイル選択

学習の実施

- 教師データ

ObjectData:
DATA: string
MKLPVRLLVLMFWIPASSSDVVMTQTPLSLPVSLGDQASISCRSSQSLVHSNGNTYLNWYLQ
KAQQSPKLLIYKVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGIYFCSQTTHVPFTFGGGT
KLEIKR

LQQPGAELVKPGASVKLSCKASGYTFTNYWIHWVKQRPGRGLEWIGRIDPNSGGTKYNEKF
KNKATLTINKPSNTAYMQLSSLTSDDSAVYYCARGYDYSYYAMDYWGQGTSVTVSS

ESGGGLVQPGGSMKLSCVASGFTFSNYWMNWVRQSPEKGLEWVAEIRLKSGYATHYAESV
KGRFTISRDDSKSSVYLQMNNLRAEDTGIYYCTRPGVPDYWGQGTTLTVSS

MEFGLSWIFLVAILKGVQCEVRLVESGGDLVEPGGSLRVSCEVSGFIFSKAWMNWVRQAPG
KGLQWVGQIKNKVDGGTTDYAAPVKGRFIISRDDSKSTVYLQMNRLKIEDTAVYYCVGNYT
GTVDYWGQGTLVTVSS

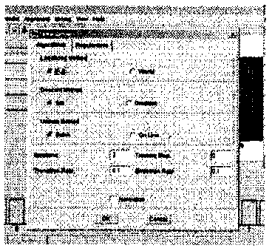
ISCKASGYTFTNYGMNWVKQAPGKGLKWMGWINTYTGEPTYADDFKGRFAFSLETSASTA
YLQINNLKNEDTATYFCARGSSYDYYAMDYWGQGTSVTVSS

LVQLQQSGPVLVKPGTSMKISCKTSGYSFTGYTMSWVRQSHGKSLEWIGLIIPSNGGTNYNQ
KFKDKASLTVDKSSSTAYMELLSLTSEDSAVYYCARFSYYGSRNYYAMDYWGQGTSVTVS
SAK

.....

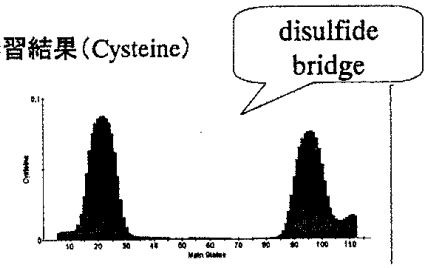
学習の実施

- アルゴリズム・パラメータ設定



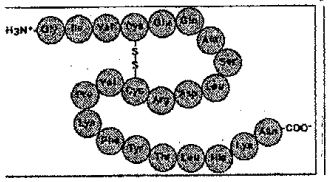
学習結果

- 学習結果(Cysteine)



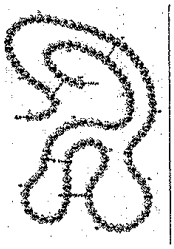
Disulfide bridge?

- 概念図



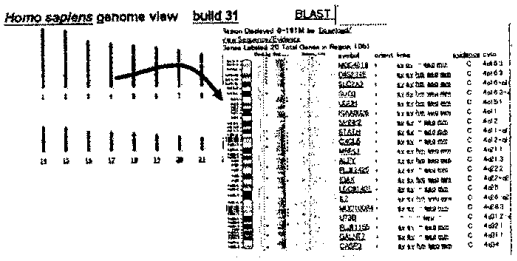
Disulfide bridge

- Lysozyme(4箇所出現)

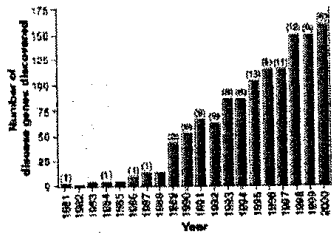


インパクト
遺伝子発見

- データベースに蓄積→遺伝子地図



インパクト
医療方面への応用



計算機科学への要請

- インタフェース
- 最適化
- 全文検索
- 圧縮

— Ugly programming instead of deep biological insights

全文検索
索引作成

- サフィックスとインデクスポイント

Text	a	b	r	a	c	a	d	a	b	r	a
Index	0	1	2	3	4	5	6	7	8	9	10

Suffix	Index
abracadabra	0
bracadabra	1
racadabra	2
acadabra	3
cadabra	4
adabra	5
dabra	6
abra	7
bra	8
ra	9
a	10

全文検索
索引作成

Suffix array (索引)

- ソート

Sorted Suffix	Index
a	10
abra	7
abracadabra	0
acadabra	3
adabra	5
bra	8
bracadabra	1
cadabra	4
dabra	6
ra	9
racadabra	2

全文検索
索引の利用

- 二分探索

"ra"

Sorted Suffix	Index
a	10
abra	7
abracadabra	0
acadabra	3
adabra	5
bra	8 ←1
bracadabra	1
cadabra	4
dabra	6 ←2
ra	9 ←3
racadabra	2

全文検索
索引作成アルゴリズム

- ブロックソートに帰着

入力文字列＋$

Suffix array

- $abcabc	3 abc$abc
0 abcabc$	0 abcabc$
1 bcabc$a	4 bc$abca
2 cabc$bc	1 bcabc$a
3 abc$abc	5 c$abcab
4 bc$abca	2 cabc$bc
5 c$abcab	

ソート

$：いずれの文字よりも小

ブロックソート

入力文字列

0 AbcABC	3 ABCAbc
1 bcABCA	0 AbcABC
2 cABCAb	4 BCAbcA
3 ABCAbc	5 CAbcAB
4 BCAbcA	1 bcABCA
5 CAbcAB	2 cABCAb

ソート

入力が特殊なソート

ブロックソートによる変換
Burrows-Wheeler 変換

- 着眼

入力文字列

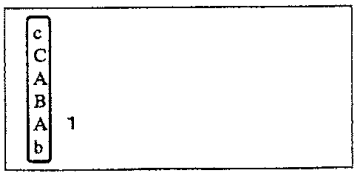
0 AbcABC	3 ABCAb c
1 bcABCA	0 AbcAB C
2 cABCAb	4 BCAbc A
3 ABCAbc	5 CAbcA B
4 BCAbcA	1 bcABC A
5 CAbcAB	2 cABCA b

ソート

最終列と1番行の場所で全情報再現可能

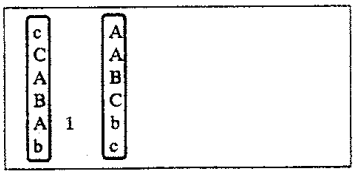
Burrows-Wheeler 変換
逆変換

- 入力



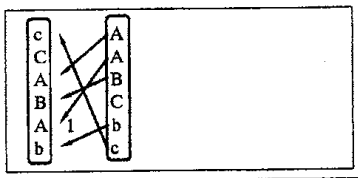
Burrows-Wheeler 変換
逆変換

- ソートによって最初の行を知る



Burrows-Wheeler 変換
逆変換

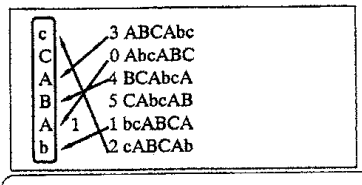
- 行番号特定



(i+1)番行の最終文字
= i番行の先頭文字

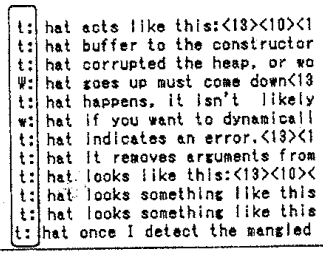
Burrows-Wheeler 変換
逆変換

- Suffix array 及び 元の文字列判明



suffix array・元の文字列判明

Burrows-Wheeler 変換
メリット



繰り返しが現れやすい→符号化効率良好

圧縮と検索の両立

ブロックソート

BWTによる高性能圧縮
＋
suffix array による高速全文検索

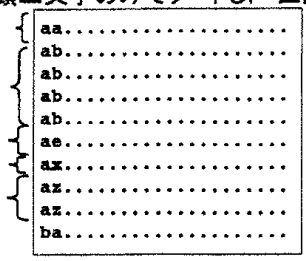
ブロックソート アルゴリズム

- quick sort
  一般アルゴリズムなので向かない
- ternary partitioning[Bentley, Sedgewick 97]
  無駄な文字列比較が少ない
- doubling algorithm
  多くの場合最速
- copy method
  対象の性質を利用
- layer method
  copy method の改良

'04田辺・小林

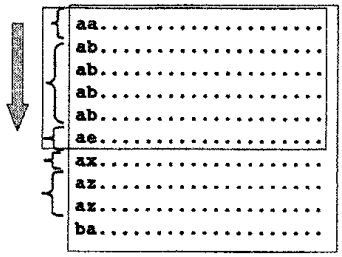
copy method 原理

- 先頭二文字のみでソートし、「区間」決定



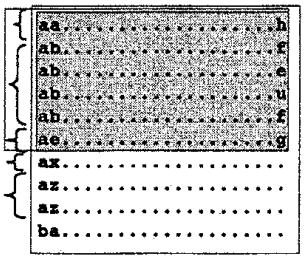
copy method 原理

- 先頭文字が同一の「区間」内を順にソート



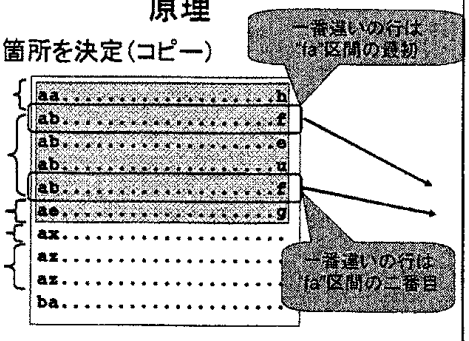
copy method 原理

- ソートされた区間内の末尾文字を調べる



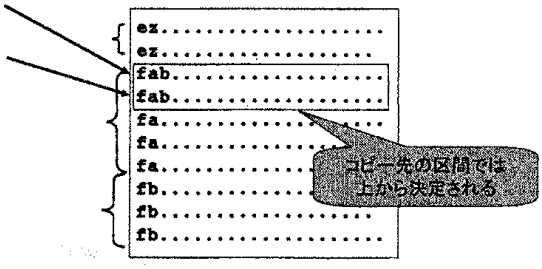
copy method 原理

- 判明箇所を決定(コピー)



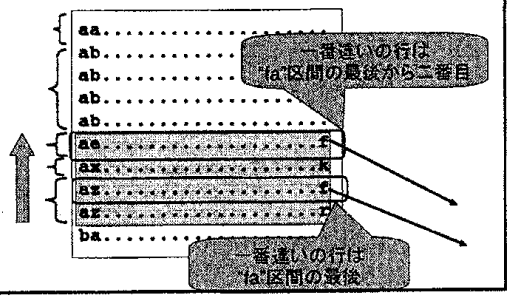
copy method 原理

- 判明箇所を決定(コピー)



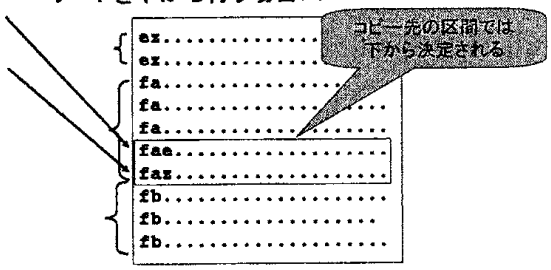
copy method
原理

- ソートを下から行う場合. .



copy method
原理

- ソートを下から行う場合. .



copy method
高速化

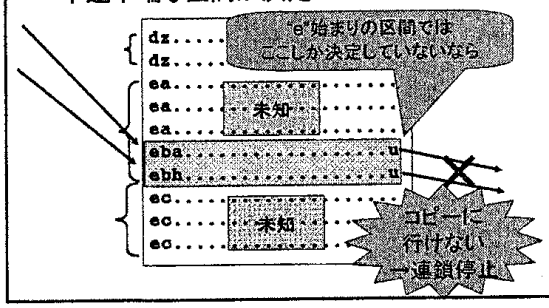
- コピー機会の増大
- ソート区間長さの減少

- ソートは小区間から
- 上下方向の使い分け

コピーの連鎖を起こしたい

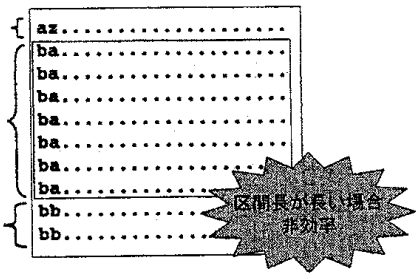
copy method
限界

- 中途半端な区間が決定..



copy method
限界

- 区間内ソート必須



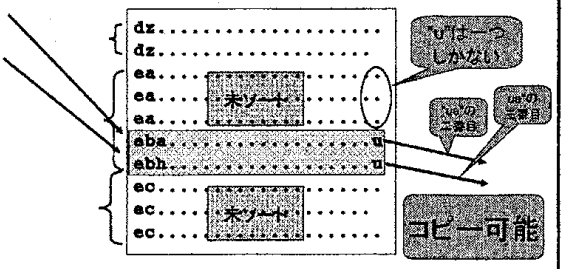
layer method
意図

- いつでもコピー可の条件確保

- コピーの連鎖を止めない

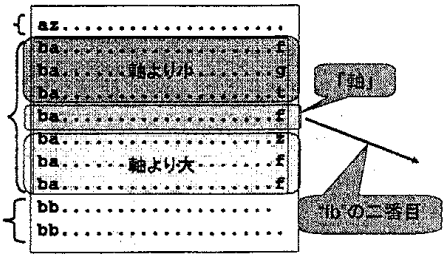
layer method 着眼

- 区間内文字数カウントがあれば..



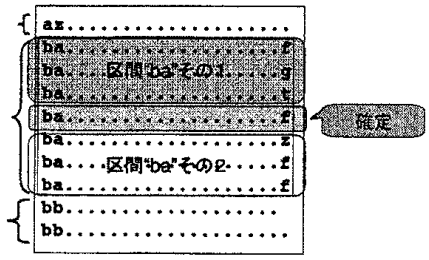
layer method 着眼

- 区間内分割ができれば軸はコピー可能



layer method 着眼

- 軸による区間分割
  →新たな軸に対する区間分割コスト削減



layer method アルゴリズム

0. 文字数カウント設定，初期区間と軸の決定
1. 軸に対する区間分割と軸のコピー
2. 文字数カウント更新
3. コピー回数＝文字数なら終了
   さもなければ
   軸をコピー先区間でも軸に設定
   1. に戻る

軸位置確定→区間分割が容易

初期区間設定以外ソート不要

layer method 実装と今後

課題:

文字数カウントの効率的更新
アルゴリズムの理論的解析

変更余地:

軸を複数行に
区間分割機会
最小化のための軸設定

参考文献

- Suffix Array 説明
  http://namazu.org/~satoru/sary/docs/suffix-array.html
- BTW解説記事
  http://www.dogma.net/markn/articles/bwt/bwt.htm
- K. Sadakane: A Modified Burrows-Wheeler Transformation for Case-insensitive Search with Application to Suffix Array Compression, To appear in Proc. of Data Compression Conference '99. (poster session) 600dpi ps.gz file (36612 bytes), PowerPoint file (65536 bytes).
  http://naomi.is.s.u-tokyo.ac.jp/~sada/papers/index.html

上記の記事の図表を使用させていただきました.